Neuron, Vol. 28, 617–625, November, 2000, Copyright ©2000 by Cell Press

# Audiovisual Integration of Letters in the Human Brain

**Tommi Raij,* Kimmo Uutela, and Riitta Hari**
Brain Research Unit
Low Temperature Laboratory
Helsinki University of Technology
P.O. Box 2200
FIN-02015-HUT
Espoo
Finland

## Summary

**Letters of the alphabet have auditory (phonemic) and visual (graphemic) qualities. To investigate the neural representations of such audiovisual objects, we recorded neuromagnetic cortical responses to auditorily, visually, and audiovisually presented single letters. The auditory and visual brain activations first converged around 225 ms after stimulus onset and then interacted predominantly in the right temporo-occipito-parietal junction (280–345 ms) and the left (380–540 ms) and right (450–535 ms) superior temporal sulci. These multisensory brain areas, playing a role in audiovisual integration of phonemes and graphemes, participate in the neural network supporting the supramodal concept of a "letter." The dynamics of these functions bring new insight into the interplay between sensory and association cortices during object recognition.**

## Introduction

Concepts are the vessels for abstract thought. Generally, a concept describes an entity in the external or internal world. Such entities typically have qualities in several sensory or motor modalities, each resulting in a different neural representation, e.g., the concept "cat" is associated with a number of visual, auditory, tactile, and olfactory properties (Damasio and Damasio, 1992).

The brain constructs multisensory interpretations of objects; spatially overlapping and simultaneous inputs from different sensory channels are merged into a unified percept (Stein and Meredith, 1993). The neural correlates of such integrative functions in humans are largely unknown. Typically, neurons in the primary sensory areas respond to stimuli in one sensory modality, whereas some neurons in the association areas (Thompson et al., 1962; Pandya and Yeterian, 1985) respond specifically to combinations of different modalities, such as audiovisual (Benevento et al., 1977) stimuli.

Sensory-specific cortices feed multisensory areas that contain unimodal patches representing different sensory modalities; multisensory neurons are often found at zones between the patches (Clemo et al., 1991; Seltzer et al., 1996; Cusick, 1997). Such an organization would first converge signals from different sensory modalities to allow them then to interact. Therefore, brain areas participating in, e.g., audiovisual integration would be expected to show signs of (1) *convergence* (both auditory and visual stimuli should activate the same region) and (2) *interaction* (the activation evoked by audiovisual stimulation should differ from the sum of unimodally presented auditory and visual activations).

Our aim was to study the human brain's audiovisual integration mechanisms for letters, i.e., for stimuli that have been previously associated through learning. For literate people, the alphabet is effortlessly transformed between the auditory and visual domains (and transmitted to the motor systems for speech and writing). Our subjects received auditory, visual, and audiovisual letters of the roman alphabet and were required to identify them, regardless of stimulus modality. Audiovisual letters included matching letters, in which the auditory and visual stimulus corresponded to each other based on previous experience, and nonmatching (randomly paired) letters. Meaningless auditory, visual, and audiovisual control stimuli were presented as well. The brain activations were detected with magnetoencephalography (MEG), which is well suited for noninvasive identification of cortical activity and its accurate temporal dynamics.

## Results

### Behavioral Results

Reaction times (RTs, finger lift latencies for target stimuli) were 505 $\pm$ 20 ms (mean $\pm$ SEM) for auditory and 520 $\pm$ 30 ms for visual letters and significantly shorter, 425 $\pm$ 15 ms ($p < 0.01$, $n = 8$, Student's two-tailed paired t tests), for audiovisual letters. The cumulative reaction time distributions further showed that RTs were faster for audiovisual letters than would have been predicted by separate processing of the auditory and visual stimuli (Raab, 1962; Miller, 1986; Schröger and Widmann, 1998). False positive or negative responses were extremely rare.

### Modality-Specific Early Activations

Figure 1 shows the grand average activations (minimum current estimates) for auditory, visual, and audiovisual letters 60–120 ms after stimulus onset. As expected, the auditory stimuli activated the supratemporal auditory cortices, and the visual stimuli activated the occipital visual areas close to midline; the audiovisual stimuli seemed to activate the areas that were activated by auditory and visual unimodal stimuli. At this early latency, activations to letters and control stimuli (data not shown) were quite similar; at later latencies, some differences were observed. To detect convergence of auditory and visual activations and audiovisual interactions, more sophisticated analysis methods were employed.

### Convergence of Auditory and Visual Activations

The auditory and visual activations converged (i.e., the activated areas overlapped) maximally in the lateral mid-

*To whom correspondence should be addressed (e-mail: tommi@neuro.hut.fi).

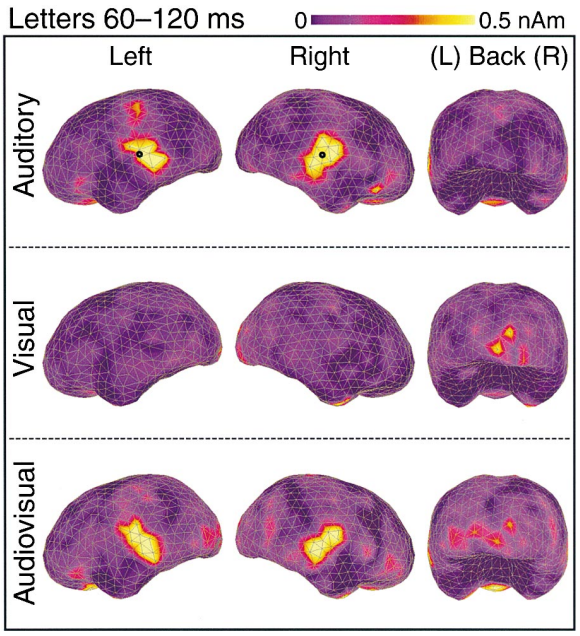


Figure 1. Early Brain Activations

Minimum current estimate (MCE) source locations for auditory, visual, and audiovisual letters from 60 to 120 ms, displayed on the surface of a triangulated standard brain. The auditory stimuli activated the supratemporal auditory cortices bilaterally, while the visual stimuli activated the occipital visual cortex. The audiovisual stimuli activated both types of sensory-specific cortices. The occipital activations are weaker than the temporal activations (the visual stimuli were small and simple). The size of the MCE color spot, projected to the surface of the boundary element model, does not only reflect the size of the activated brain area (a larger area producing a larger spot); it also depends on the depth of the activation (a deeper source is reflected as a larger spot on the surface of the brain model) and on the strength of the activation (a stronger activation results in a brighter and larger spot). The black dots in the supratemporal cortices bilaterally show the source locations for the auditory 100 ms responses (the corresponding Talairach coordinates are listed in Table 1).

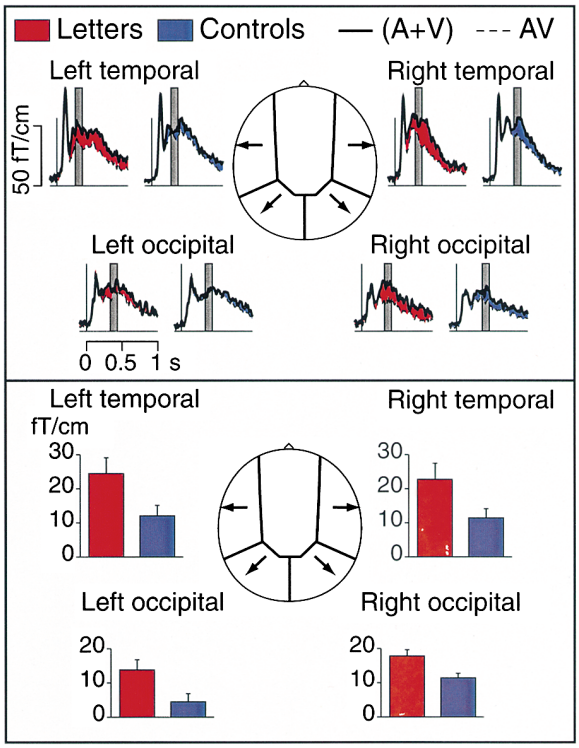


Figure 2. A + V versus AV Responses

(Top) Grand average response waveforms. The difference between the traces reflects the audiovisual interaction.

(Bottom) The ([A + V] – AV) subtraction waveform amplitudes (bars) across subjects (data from individual subjects), separately for matching audiovisual letters (AVLm) and control stimuli. For this comparison, the signals from the two orthogonal sensors at each of the 61 measurement locations were combined (vector summing, see Experimental Procedures).

temporal areas. Convergence areas, time courses, and strengths were quite similar for letters and controls. Convergence is characterized further in conjunction with sources of audiovisual interaction.

### Audiovisual Interaction: Signals

Figure 2 compares the sum of responses to auditory and visual stimuli (A + V) with responses to audiovisual stimuli (AV); the difference reflects audiovisual interaction. The upper panel shows grand average A + V and AV responses over four brain areas (the left and right temporal and occipital areas) showing the largest interaction effect, separately for matching audiovisual letters (AVLm) and control stimuli. In the great majority of cases, the effect was clearly suppressive (AV $<$ A + V), suggesting that the simultaneous A and V stimuli inhibit each other. The lower panel shows the mean $\pm$ SEM differences ([A + V] $-$ AV within a 100 ms time window centered at the latency of the difference maximum). The letters showed a significantly stronger interaction than controls, both across the four areas ($p < 0.001$, $n = 8$, Student's two-tailed paired t-test, collapsed within subjects) and within areas ($p < 0.01$, $n = 8$, for each area separately).

The interaction was strongest at 345 $\pm$ 20 ms (mean $\pm$ SEM) for letters and at 375 $\pm$ 20 ms for control stimuli, without significant latency differences between the four areas. The strength of interaction significantly exceeded the noise level from 275 $\pm$ 15 to 495 $\pm$ 40 ms for letters and from 310 $\pm$ 20 to 435 $\pm$ 25 ms for control stimuli.

Interaction for nonmatching audiovisual letters (AVLnm) was also suppressive and significantly stronger than for control stimuli ($p < 0.001$ across the areas, $n = 8$); the difference was significant in the bilateral temporal and right occipital areas ($p = 0.002$–$0.024$, $n = 8$). For these signals, the strengths of interaction did not differ between matching and nonmatching audiovisual letters. The effect was maximal at 370 $\pm$ 30 ms and significantly above noise level from 275 $\pm$ 20 to 475 $\pm$ 40 ms.

The above values were picked from channels showing the maximum interaction effect for letters. In channels showing maximal interaction for control stimuli, the interaction was typically about equally strong for letters and controls. Thus, interaction could occur for both letters and control stimuli, but some areas showed significantly stronger interaction for letters.

In addition to the suppressive interaction described above, two subjects showed clear potentiation at some

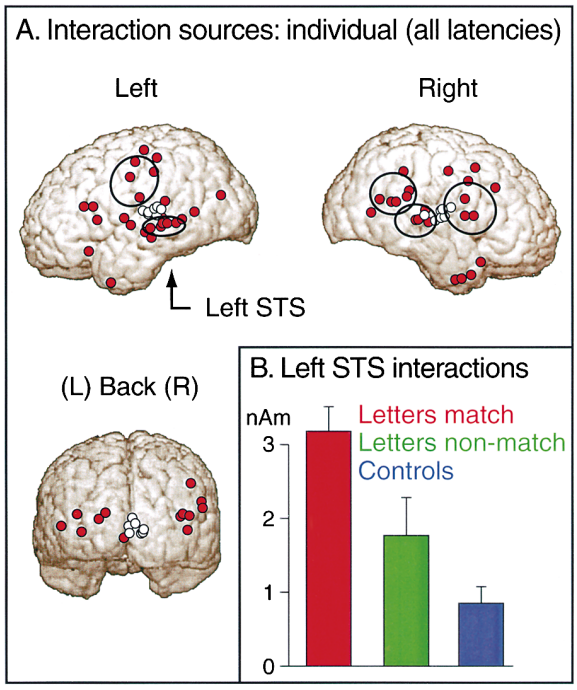


Figure 3. Individual Sources of Interaction for Letters

(A) Interaction sources from all subjects at all latencies on a standard brain. The red dots indicate the areas showing audiovisual interaction, while the white dots show the auditory and visual sensory projection cortices. The black circles outline the main grand average interaction source areas (Figure 4A, upper row).

(B) Interaction (mean ± SEM) across subjects in left STS, separately for matching letters, nonmatching letters, and control stimuli.

latencies over few areas; such cases were, however, too few to allow meaningful statistical comparisons. Response potentiations were thus not characterized further.

### Audiovisual Interaction: Source Activations

Figure 3A shows the individual interaction areas for letters at all latencies. The sources (red dots) from individual subjects are projected on the surface of a standard brain. The bilateral supratemporal (ST) auditory cortices (generators of the auditory 100 ms response) and the visual calcarine cortex (generators of the visual 90 ms response) are shown for comparison (white dots). Interaction was most consistently observed in five brain areas. In the left hemisphere, all eight subjects showed clear interaction in the superior temporal lobe (seven in the superior temporal sulcus (STS) and one in posterior ST cortex) and four subjects in the frontoparietal region (LFP). In the right hemisphere, the main interactions occurred in the frontal cortex (RF, six subjects), the temporo-occipito-parietal junction (RTOP, seven subjects), and in the STS (four subjects).

Figure 3B compares the interaction in the left STS sources across stimulus categories (data from individual subjects). Interaction was strongest for matching letters at 390 ± 50 ms (mean ± SEM); the interaction was 47% weaker for nonmatching letters ($p = 0.02$, $n = 8$) and 73% weaker for control stimuli ($p < 0.001$, $n = 8$) in the 100 ms time window centered at the interaction maximum for matching letters.

Figure 4A shows the grand average areas for interaction at 380–540 ms, separately for matching letters (upper) and control stimuli (lower). Table 1 lists the Talairach coordinates and the interaction latencies for these source areas, along with the coordinates of the auditory and visual projection cortices. Again, audiovisual interaction was prominent in five brain areas. The LFP and RF regions showed interaction starting at about 160 ms (earlier than the time window presented here), without clear differences between letters and controls. Interaction in the RTOP starting at 280 ms was followed by interaction at 380 ms in the left and 70 ms later in the STS; these three areas showed stronger interaction for letters than for control stimuli.

Figure 4B shows the grand average time courses of left STS activation for auditory, visual, and audiovisual stimulation, separately for matching letters versus nonmatching letters versus control stimuli. The first gray belt (C) shows the time span when the auditory and visual activations converged in the left STS. Convergence reached its maximum in STS at 200 ms (above 2/3 of maximum at 125–445 ms); convergence time span was quite similar across the above five brain regions (maximum at 225 ± 10 ms). In all these areas, convergence was similar for letters and controls. The later gray belt (IA) shows the time span when audiovisual interaction (lowest panel), evident as smaller responses to audiovisual than auditory stimuli (arrows), occurred in left STS (maximal at 465 ms, above 2/3 of maximum at 380–540 ms); interaction was 61% weaker for controls than for matching letters in the time period 465 ± 50 ms (peak latency ± 50 ms). RTOP showed 43% and right STS 67% weaker interaction for controls than for matching letters.

Clearly different interactions for matching than nonmatching audiovisual letters were observed in both left and right STS; the effect was 57% weaker in the left STS and 58% weaker in the right STS for nonmatching letters (time windows 465 ± 50 and 495 ± 50 ms, respectively). In RTOP, the interaction was strongest for nonmatching (340 ± 50 ms) and fairly similar for matching letters; the effect was 54%/43% weaker for controls than for nonmatching and matching letters, respectively. All the reported differences between categories clearly exceed the noise level, being 4.2 ± 0.4 times stronger than activity during the prestimulus baseline.

## Discussion

### Cortical Network Supporting Audiovisual Integration

In the present study, we were able to identify the multisensory cortical network that combines auditory (phonemic) and visual (graphemic) aspects of letters of the alphabet and to determine the time courses of the associated events. For audiovisual stimuli, the sensory-specific auditory and visual projection areas were first activated strongly at 60–120 ms. These activations were apparently forwarded to multisensory areas that around 225 ms received maximal input from both sensory modalities as a sign of convergence. For matching letters,

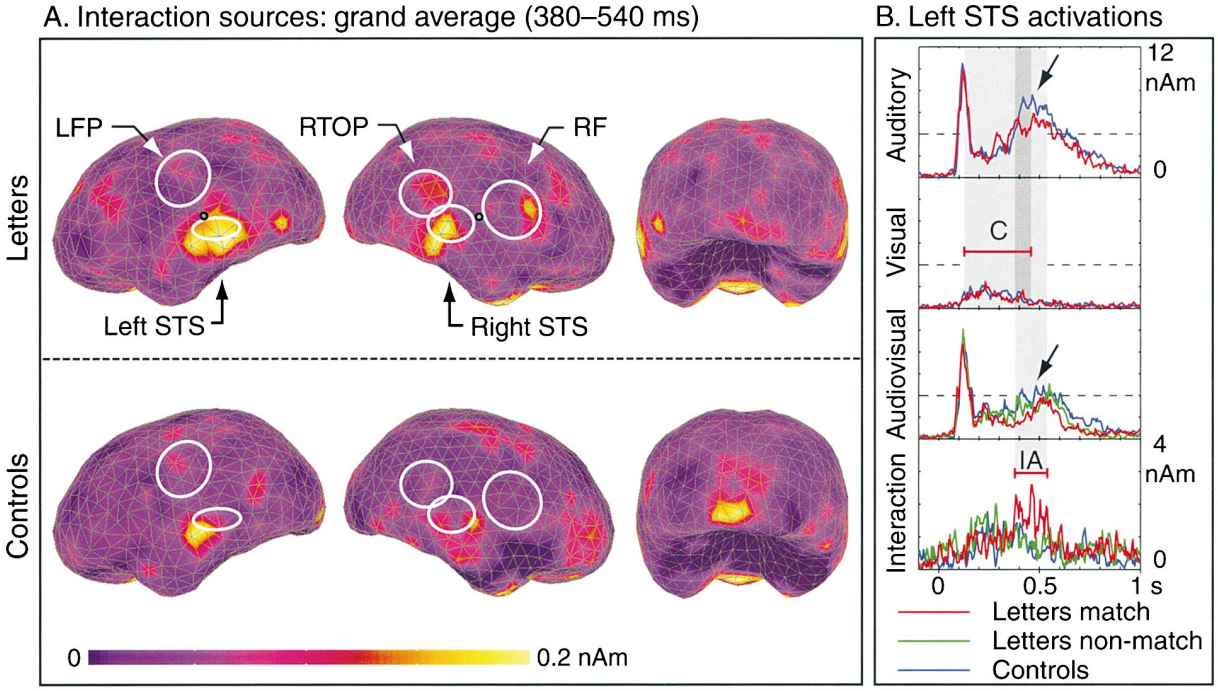


Figure 4. Grand Average Interaction Sources and Time Courses of Left STS Activations

(A) The interaction sources are shown separately for letters and control stimuli. The effect is shown from 380 to 540 ms, when the bilateral STS sources showed strongest interaction. The main interaction areas are marked with white circles. Interaction in LFP, RF, and, partially, even in RTOP occurred earlier than the illustrated time window; consequently, these sources are not optimally visible. Some discrepancy between the circles and the color maps stems from the different procedure in projecting the sources to the brain surface (the circles directly toward the viewing direction and the color maps along the radius of the conductor model).

(B) Grand average activation time courses of the left STS interaction source. The three upper panels show the time courses for auditory, visual, and audiovisual stimulation, separately for letters (audiovisual separately for matching and nonmatching letters) and control stimuli. The arrows point at the activation that was dampened for audiovisual as compared with auditory stimuli. The gray shadings highlight the time windows when convergence of the auditory and visual activations (C) and audiovisual interaction (IA) were maximal. For interaction (lowest curves), the differences between matching and nonmatching letters were 4.5 times and between matching letters and control stimuli 4.9 times stronger (time window, 465 ± 50 ms) than the prestimulus noise level. Amplitude scales, 0–12 nAm for letters (horizontal dashed lines at 4 nAm) and 0–4 nAm for interaction; the time scale is from –100 to 1000 ms.

in which the auditory and visual stimuli had been associated through extensive previous learning, we observed a suppressive interaction around 380–540 ms. For control stimuli and nonmatching letters, the interaction was significantly weaker. Thus, as a result of convergence and interaction of the auditory and visual activations, the phoneme and the grapheme were integrated.

We consider, for several reasons, that these observations really reflect multisensory integration in the human brain. (1) The experimental design required the subjects to relate the auditory and visual letters to each other. (2) The reaction times were faster for audiovisual than for unimodal stimuli. This phenomenon could result from two different mechanisms. The audiovisual stimuli might be processed separately in the auditory and visual domains, and the quicker of the two processes could initi-

Table 1. Talairach Coordinates and Latencies of the Sources

| Sensory Projection Cortices (60–120 ms) | x | y | z | | |
|---|---|---|---|---|---|
| Left auditory cortex | −53 | −25 | +10 | | |
| Right auditory cortex | +55 | −14 | +12 | | |
| Visual primary cortex | +7 | −81 | +6 | | |
| | | | | Interaction Latencies | |
| Interaction Sources | x | y | z | Peak | (Above 2/3 of Peak) |
| Left frontoparietal (LFP) | −46 | −17 | +35 | 245 ms | (155–250 ms) |
| Left superior temporal sulcus (LSTS) | −53 | −31 | 0 | 465 ms | (380–540 ms) |
| Right frontal (RF) | +42 | +4 | +21 | 305 ms | (160–470 ms) |
| Right temporo-occipito-parietal (RTOP) | +45 | −49 | +20 | 340 ms | (280–345 ms) |
| Right superior temporal sulcus (RSTS) | +48 | −31 | +6 | 495 ms | (450–535 ms) |

ate the motor response ("race model"; Raab, 1962). Alternatively, the speeded RTs could result from combined processing of the auditory and visual stimuli ("coactivation model"; Miller, 1986). In accordance with previous results (Miller, 1986; Schröger and Widmann, 1998), the cumulative RT distributions revealed that the speeded RTs reflected audiovisual integration not race model–like competition. (3) The brain activations triggered by auditory and visual stimuli converged. (4) The sum of unimodal activations differed from audiovisual activation, implying multisensory interaction. (5) In some brain areas, the interactions differed between matching letters, nonmatching letters, and control stimuli, showing that the type and combination of stimuli were also important for the interaction.

The audiovisual convergence, clearest in the temporal areas bilaterally, was remarkably similar for letters and controls, suggesting that the sensory-specific cortices do not necessarily gate the access to multisensory areas, so that even unrecognized stimuli can proceed further. Consequently, any two stimuli can be associated through learning, even if the relation between them is arbitrary, as for letters.

Audiovisual interaction was prominent in five brain areas. The LFP and the RF regions showed interaction quite early and did not differentiate between letters and control stimuli. The interaction in the RTOP and the left and right STS occurred later and was stronger for letters than for controls. This finding suggests that the overlearned association between phonemes and graphemes has resulted in an organizational change in these brain areas. In the following, we discuss separately the main areas participating in the neural network where the audiovisual interaction occurred.

The left posterior STS, a part of Wernicke's area, showed prominent audiovisual integration of letters in all eight subjects. The primate STS contains auditory, visual, and association areas (Barnes and Pandya, 1992; Seltzer et al., 1996; Cusick, 1997). In humans, the left STS contains critical areas for comprehension of both spoken and written words (Ojemann et al., 1989; Démonet et al., 1992; Howard et al., 1992; Fiez et al., 1995; Price et al., 1996; Abdullaev and Posner, 1998), although regions that are activated by both speech and text can vary with word category (Damasio et al., 1996; Martin et al., 1996) and experimental task (Price et al., 1997; Chee et al., 1999). The left STS has also been implicated in auditory processing of visually presented letters (Sergent et al., 1992) and in visual imagery of auditorily presented letters (Raij, 1999). The current study strongly supports the role of the left STS in audiovisual encoding and transformation of single letters. The right STS has been implied in reading words and nonwords (Paulesu et al., 2000), and it could play a similar functional role as the left STS. However, as the audiovisual interaction started about 70 ms later in the right than in the left STS, the right STS signals could also reflect activation through bilaterally symmetrical callosal connections. The STS cortices of both hemispheres showed clearly stronger interaction for matching than nonmatching letters, suggesting that these areas are mainly responsible for the audiovisual integration process.

The RTOP showed clear interactions for letters, but it did not distinguish between matching and nonmatching audiovisual letters. A study of patients with RTOP lesions has suggested for this area a perceptual classification function (Warrington and Taylor, 1973), which is critically important for feature analysis of letters, as their physical properties in natural speech and handwriting vary widely. RTOP is also activated during phonological (but not semantic) processing of visually presented words, suggesting an audiovisual conversion function (Price et al., 1997). The timing of this interaction (onset 100 ms prior to left STS) would be consistent with both functions.

The location of the LFP source in the Rolandic region would agree with somatomotor activation. However, the subjects only responded to targets and with the hand (left) ipsilateral to the Rolandic activation. The source volume extended anteriorily to areas that have been associated with audiovisual attention (frontal eye fields) and audiovisual–motor integration (Bodner et al., 1996; Paus, 1996; Iacoboni et al., 1998) and a variety of language functions (Ojemann, 1992), including semantic processing of both auditorily and visually presented words (Chee et al., 1999). Similarly as in the current study, this region has been shown to be activated quite early during visual imagery of single letters (Raij, 1999). The time course and reactivity of LFP (earliest interaction, no distinction between letters and controls) would suggest a rather general function related to audiovisual attention.

### Intermodal Potentiation versus Suppression

The brain associates across senses stimuli that might arise from the same origin (for reviews, see Stein and Meredith, 1993; Stein, 1998). For example, stimuli that occur in the same spatial location (spatial coincidence) are merged, provided that they occur more or less simultaneously (temporal coincidence). When this happens, subcortical structures (especially the superior colliculus) interact with multisensory cortex, evidently to produce orientation-related motor acts (Stein et al., 1988; Wallace and Stein, 1994; Peck et al., 1995; Wallace et al., 1996). The rules governing multisensory integration appear quite similar in subcortical and cortical structures, though some differences exist (Stein and Wallace, 1996). Typically, multisensory neurons show potentiation to spatiotemporally coinciding stimuli, so that the response to a multisensory stimulus can be even many times stronger than the corresponding unimodal activations. Multisensory suppression is also known to occur, mainly when the stimuli do not coincide across senses spatially and/or temporally (Kadunce et al., 1997).

We observed, especially in the left STS, clear suppression of the audiovisual activations, compared with auditory activations (Figure 4B, arrows). This MEG signal decrease cannot be explained by cancellation of currents in the STS region, as the source currents were similarly directed (downward) during unimodal (auditory and visual) activations. Thus, the suppression most likely reflects neuronal level interactions. How does this suppression then relate to the animal data cited above and to previous human recordings?

First, the earliest multisensory studies mainly used behaviorally irrelevant stimuli such as flashes of light and clicks of sound. More ecologically valid stimuli have

recently been introduced, with the auditory and visual stimuli coinciding in space and time (e.g., Sams et al., 1991; Calvert et al., 2000). However, such stimuli also clearly differ from the present stimuli, which were entirely culture-based artifacts (letters of the alphabet), where the auditory (phonemic) and visual (graphemic) stimuli have, through life-long learning, been associated with each other according to totally arbitrary rules. Learning can lead to suppression of cortical responses; for example, perceptual/priming visual learning is associated with decreased signal amplitudes in electrophysiological, PET, and fMRI recordings (Kok and de Jong, 1980; Raichle et al., 1994; Büchel et al., 1999; for recent reviews, see Desimone, 1996; Wiggs and Martin, 1998). This type of learning apparently leads to optimization of activation in the local network, which can result in improved recognition of the stimulus in noisy conditions (Dosher and Lu, 1998; Gold et al., 1999). The process could be compared to sharpening of neuronal tuning, resulting in suppressed responses (Hurlbert, 2000). Accordingly, when the simultaneously presented grapheme and phoneme "match" with each other according to previous experience (overlearned situation), the responses in the local neural network in STS could be relatively suppressed. The stronger signal amplitudes for nonmatching audiovisual letters and audiovisual control stimuli, in which the audiovisual combination is novel, might reflect suboptimal tuning in the local network.

Second, it is to be noted that, until now, multisensory potentiation has mainly been shown at the level of single neurons. Noninvasive electroencephalographic (EEG) and MEG recordings pick up synchronous (mass) activity of thousands of neurons, and previous EEG and MEG multisensory studies have almost exclusively shown suppressive not potentiative audiovisual interactions (Morrell, 1968; Davis et al., 1972; Squires et al., 1977; Busch et al., 1989; Schröger and Widmann, 1998; Giard and Peronnet, 1999). It thus seems that the net audiovisual interaction effect can be suppressive, while some neurons show multisensory potentiation.

Letters are abstract concepts in the sense that they are effortlessly transformed between the auditory and visual domains, even in the absence of spatial and temporal coincidence between the phonemes and graphemes. During learning of the alphabet, spatiotemporal coincidence is, however, apparently required; when you were taught letters of the alphabet, the school teacher probably tapped at a mystical figure at the blackboard while producing the "corresponding" sound. In learning to associate a given grapheme with a certain phoneme, both potentiative and suppressive processes apparently take place in the local network (reviewed in Bear, 1996). Our results suggest that, at the learning phase, one would expect a weaker net audiovisual suppression than in the fully learned situation. Thus, learned abstract audiovisual associations are apparently reflected as suppression in EEG/MEG recordings of the association cortex.

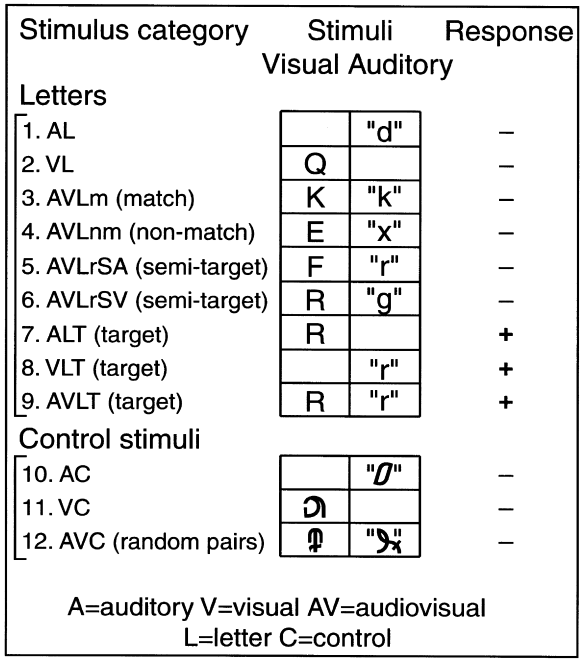


Figure 5. Stimuli and Tasks

The letter "R" (auditory, visual, or audiovisual) is the current target. The stimuli were presented in a randomly ordered sequence, where a single stimulus could represent any of the 12 categories. The required response to the target was a left index finger lift. We used the letter names as auditory stimuli, but, because in the Finnish language a letter is pronounced always similarly, regardless of other surrounding letters, these did not largely differ from the associated phonemes. However, the neural networks converting graphemes to phonemes might be partially differently organized in different languages. For example, Italians reading aloud visually presented words (and nonwords) activate the left superior temporal gyrus more strongly than English readers, probably because Italian (like Finnish) has a very consistent grapheme–phoneme relation, whereas, in English, a letter can correspond to many phonemic expressions (Paulesu et al., 2000).

### Neuronal Representation of Concepts and Multisensory Binding

The current study offers some insight into the neural representations of abstract concepts in general. Unimodal representations in sensory-specific cortices have been suggested to communicate through multisensory nodes (Mesulam, 1998). Our results agree with such a view by showing that integration occurs mainly in areas other than the sensory-specific auditory or visual cortices; recognition and recall of multisensory aspects of concepts should, thus, critically depend on proper functioning of the multisensory nodes. Damage to different parts of the network supporting a concept apparently results in different types of functional deficits (McCarthy and Warrington, 1988; Damasio et al., 1996; Martin et al., 1996). Although we did not find audiovisual interaction in sensory-specific cortices, the division between unimodal and multisensory areas is not absolute; sensory projection cortices can, under certain conditions, receive modulating input from other modalities (Bental et al., 1968; Sams et al., 1991; Yaka et al., 1999).

One of the greatest challenges in neuroscience is to understand how different parts of the brain, receiving information about the external world through different sensory channels, communicate to produce a holistic internal representation of a given object. This "binding problem" has mainly been studied within the visual sys-

tem and by means of the brain's oscillatory signals (Eckhorn et al., 1988; Livingstone and Hubel, 1988; Gray and Singer, 1989; Roelfsema et al., 1996). The problem is also multisensory, and the present study shows clear learning-based binding across modalities. Clarifying the neural basis of multisensory integration should greatly enhance understanding the brain implementation of binding in general.

## Experimental Procedures

### Subjects and Stimuli

Nine healthy literate adults (age 22–32 years, five males, eight right handed) were presented with a sequence consisting of auditory, visual, and audiovisual (simultaneous auditory and visual) stimuli. The auditory stimuli were digital recordings of 20 phonemic expressions of the Finnish language, representing single letters (names of letters ACDEGHIJKLNOPQRSTUVY, duration 300 ± 10 ms), and of 20 different auditory control stimuli that were processed from the letter stimuli to become unpronounceable and unrecognizable as letters but were of the same duration and contained the same general amplitude envelope and carrier frequency. The monophonic sounds were delivered to the subjects binaurally through plastic tubes and earpieces.

The visual stimuli were capital letters corresponding to the 20 auditory letters (ACDEGHIJKLNOPQRSTUVY, duration 255 ms) and 20 control nonletter stimuli prepared by decomposing the letters and rotating and shifting individual parts of them (no symbols carrying a semantic meaning were allowed). The visual stimuli covered 4° of the central visual field and were presented on a rectangular white background, projected into the measurement chamber with a data projector.

Figure 5 shows the 12 different stimulus categories. All 12 different types of stimuli were presented in a single, randomly ordered sequence, once every 1.5 s; evoked responses were averaged separately for each category. All categories were equiprobable, except categories five and six, which were half as probable as any other single category. At least three different stimuli occurred between successive presentations of the same stimulus. For audiovisual control stimuli, any auditory control could appear randomly with any of the visual controls, and the pairs were not fixed.

The subject's task was to lift the left index finger as quickly and accurately as possible to a target letter. The target probability was evenly distributed across all letters. The target was changed randomly (on average, every 50 stimuli) with a preceding audiovisual warning stimulus, followed by audiovisual presentation of the new target. For audiovisual targets, the same letter was presented auditorily and visually, whereas the subject was instructed not to lift the finger for nonmatching audiovisually presented letters where one stimulus was the target while the other was not ("semitargets"). The task thus required the subjects to relate the auditory and visual letters to each other.

The recordings were carried out during two identical 30 min sessions on separate days. All necessary instructions were given immediately before the measurement.

### Recordings and Data Analysis

Cerebral magnetic signals were recorded with a whole-scalp 122 channel planar SQUID (superconducting quantum interference device) magnetometer (Neuromag-122) (Ahonen et al., 1993) in a magnetically shielded room. The instrument measures two orthogonal tangential derivatives of the magnetic field at 61 measurement sites, giving the largest signal just above a dipolar source (for a review of MEG, see Hämäläinen et al., 1993). The signals were band-pass filtered at 0.03–100 Hz and digitized at 397 Hz.

The signals were averaged online, with an analysis period extending from 100 ms prestimulus to 1500 ms poststimulus. Vertical and horizontal electrooculograms (passband 0.3–100 Hz) were recorded from electrodes above and below the left eye and lateral to the eyes, and epochs contaminated by eyeblinks or eye movements (signals exceeding ±150 μV) were automatically discarded from the averages. During offline analysis of the signals, the averaged signals were digitally low-pass filtered at 40 Hz, and response amplitudes were measured with respect to a 100 ms prestimulus baseline.

Anatomical data for all subjects were obtained from 1.5 T magnetic resonance images (MRIs); one subject was excluded from MEG analysis due to an arachnoideal cyst in the MRI. To align MEG and MRI data, the position of the head with respect to the sensor array was calculated from magnetic signals produced by currents fed into three small coils attached to the scalp (Ahonen et al., 1993). The position of the coils with respect to the nasion and the two preauricular points was measured with a 3D digitizer.

The responses from the two recording sessions were found to be highly replicable within subjects; thus, they were averaged to further increase the S/N ratio.

Averaged responses were compared across categories. To exclude responses associated with immediate motor processes, only responses to nontargets were considered. Two main comparisons were made to reveal audiovisual interaction effects: (1) (AL + VL) – AVLm, the sum of responses to auditory and visual letters minus responses to audiovisual letters (matching pairs); (2) (AC + VC) – AVC, the sum of responses to auditory and visual controls minus responses to audiovisual controls (random pairs). A third comparison, (3) (AL + VL) – AVLnm, the sum of responses to auditory and visual letters minus responses to nonmatching audiovisual letters (random pairs), was made as well.

The response strengths were compared after calculating the vector sums for each orthogonal channel pair at the 61 measurement locations:

$$\sqrt{(\delta B_z/\delta x)^2 + (\delta B_z/\delta y)^2}$$

where $B_z$ is the measured magnetic flux component.

Vector sums simplify the analysis of evoked responses when the orientation of the source current varies strongly with only small accompanying changes in location of the generating current, as often occurs in highly convoluted cortical areas. The (A + V) sum was calculated before vector summing; otherwise, negative A and V signal values could have produced uncontrolled results. The responses and the source estimates are slightly biased toward stronger A + V than AV activity due to measurement noise, but this is negligible, as can be clearly seen from the diminutive difference during the baseline period (Figure 2). The bias does not affect comparisons between letters and control stimuli because noise is similar for both.

Visual inspection of the evoked responses clearly suggested that the largest interaction effects were suppressive (i.e., the sum of responses for unimodal A and V stimulation was larger than the response for AV stimulation). Thus, for each subject, the MEG channel (vector sum of a sensor pair) showing the maximum interaction for matching letters (AVLm) was identified, and the peak latencies and mean amplitudes within ±50 ms from the peak latency were measured separately for four areas. For control stimuli (and nonmatching letters), the interaction effects were measured from the same channel to ensure that the signals were generated in about the same brain locations. Finally, the time windows when the interaction effects were >2 SD above the prestimulus noise level were measured. To ascertain that similar response components were compared, the time windows for comparisons 2 and 3 were required to overlap at least partially with the time window of comparison 1; in the great majority of cases, the comparison epochs overlapped anyway. For the interaction effect offsets, latencies before 1000 ms were considered.

### Source Estimation

To estimate the neural currents from the MEG data, we used the minimum current estimate (MCE) (Uutela et al., 1999), an implementation of the minimum *l*1-norm estimate (Matsuura and Okabe, 1995). MCE explains the measured signals with a current distribution that has the smallest sum of current amplitudes. The estimate is calculated separately for each time sample. The source area was individually restricted into the subject's brain, but the regions in the middle of the brain were neglected because such currents produce only weak magnetic fields outside the head. The cerebellum was also neglected because preliminary analysis did not indicate strong signals originating from it. As sources in the basal brain surface are

particularly sensitive to artifacts (arising from, e.g., heart and eye), some activations in temporal poles were not characterized further. To eliminate any low-frequency measurement noise, the linear trends between baselines of two successive stimuli were removed.

We then searched for areas showing significant convergence and calculated (separately for each subject, time point, and location) the MCE of auditorily and visually evoked activity and selected the smaller one (minimum). The current orientations were discarded. Because of the spatial resolution of the method, we smoothed the MCE spatially using Gaussian kernel with 1 cm width. Further, as in behavioral experiments auditory and visual stimuli can interact at least within a 100 ms time window (McGrath and Summerfield, 1985; Aunon and Keirn, 1990), we low-pass filtered the MCE temporally using a Gaussian kernel with ±50 ms width. Our claims about convergence are thus spatially accurate within 1 cm and temporally within ±50 ms.

The interaction effect was studied by calculating the MCE of the linear combination of measured responses, (A + V) − AV. The resulting MCEs and amplitude values show the absolute (rectified) interaction strengths. Thus, the suppressive and potentiative interactions can reflect equally in the source strengths.

The brain volumes showing the strongest MCEs were selected manually, with the center and extent automatically adjusted to the maximal activity. The time course of the activity within the selected volume was then calculated as a spatially weighted average of the estimate; the weight was maximal in the center of the volume and decayed radially with the form of a three-dimensional Gaussian kernel.

The estimates were studied both as a grand average across subjects and separately for each subject. For grand averages, the individual estimates were first spatially aligned with a piecewise linear transformation based on the locations of anterior and posterior commissures and the size of the brain (Talairach and Tournoux, 1988). To visualize the estimates, the activity was projected on the surface of a standard brain (Roland and Zilles, 1996) and color coded. As the activations from the left-handed subject did not differ in any major way from other subjects, his data were included in the statistical comparisons.

## Acknowledgments

This study has been financially supported by the Academy of Finland, the Sigrid Jusélius Foundation, and the Ministry of Education. The MRIs were recorded at the Department of Radiology of the Helsinki University Central Hospital. We thank G. Curio for useful discussions, M. Seppä for advice in brain surface model construction, and M. Illman for technical assistance.

Received February 24, 2000; revised September 14, 2000.

## References

Abdullaev, Y.G., and Posner, M.I. (1998). Event-related brain potential imaging of semantic encoding during processing single words. Neuroimage *7*, 1–13.

Ahonen, A.I., Hämäläinen, M.S., Kajola, M.J., Knuutila, J.E.T., Laine, P.P., Lounasmaa, O.V., Parkkonen, L.T., Simola, J.T., and Tesche, C.D. (1993). 122-channel SQUID instrument for investigating the magnetic signals from the human brain. Phys. Scripta *T49*, 198–205.

Aunon, J.I., and Keirn, Z.A. (1990). On intersensory evoked potentials. Biomed. Sci. Instrum. *26*, 33–39.

Barnes, C.L., and Pandya, D.L. (1992). Efferent cortical connections of multimodal cortex of the superior temporal sulcus in the rhesus monkey. J. Compar. Neurol. *318*, 222–244.

Bear, M.F. (1996). A synaptic basis for memory storage in the cerebral cortex. Proc. Natl. Acad. Sci. USA *93*, 13453–13459.

Benevento, L.A., Fallon, J., Davis, B.J., and Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and orbital cortex of the macaque monkey. Exp. Neurol. *57*, 849–872.

Bental, E., Dafny, N., and Feldman, S. (1968). Convergence of auditory and visual stimuli on single cells in the primary visual cortex of unanesthetized unrestrained cats. Exp. Neurol. *20*, 341–351.

Bodner, M., Kroger, J., and Fuster, J.M. (1996). Auditory memory cells in dorsolateral prefrontal cortex. Neuroreport *7*, 1905–1908.

Büchel, C., Coull, J.T., and Friston, K.J. (1999). The predictive value of changes in effective connectivity for human learning. Science *283*, 1538–1541.

Busch, C., Wilson, G., Orr, C., and Papanicolaou, A. (1989). Crossmodal interactions of auditory stimulus presentation on the visual evoked magnetic response. In Advances in Biomagnetism, S.J. Williamson, M. Hoke, G. Stroink, and M. Kotani, eds. (New York: Plenum Press), pp. 221–224.

Calvert, G.A., Campbell, R., and Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr. Biol. *10*, 649–657.

Chee, M.W.L., O'Craven, K.M., Bergida, R., Rosen, B.R., and Savoy, R.L. (1999). Auditory and visual word processing studies with fMRI. Hum. Brain Mapp. *7*, 15–28.

Clemo, H.R., Meredith, M.A., Wallace, M., and Stein, B. (1991). Is the cortex of cat anterior ectosylvian sulcus a polysensory area? Soc. Neurosci. Abstr. *17*, 1585.

Cusick, C.G. (1997). The superior temporal polysensory region in monkeys. In Cerebral Cortex: Extrastriate Cortex in Primates,Volume 12, K.S. Rockland, J.H. Kaas, and A. Peters, eds. (New York: Plenum Press), pp. 435–468.

Damasio, A.R., and Damasio, H. (1992). Brain and language. Sci. Am. *267*, 87–95.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., and Damasio, A.R. (1996). A neural basis for lexical retrieval. Nature *380*, 499–505.

Davis, H., Osterhammel, P.A., Wier, C.C., and Gjerdingen, D.B. (1972). Slow vertex potentials: interactions among auditory, tactile, electric and visual stimuli. Electroencephalogr. Clin. Neurophysiol. *33*, 537–545.

Démonet, J., Chollet, R., Ramsay, S., Cardebat, D., Nespoulous, J., Wise, R., Rascol, A., and Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. Brain *115*, 1753–1768.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. Proc. Natl. Acad. Sci. USA *93*, 13494–13499.

Dosher, B.A., and Lu, Z.L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. Proc. Natl. Acad. Sci. USA *95*, 13988–13993.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., and Reitböck, H.J. (1988). Coherent oscillations: a mechanism for feature linking in the visual cortex? Biol. Cybern. *60*, 121–130.

Fiez, J.A., Raichle, M.E., Balota, D.A., Tallal, P., and Petersen, S.E. (1995). PET activation of posterior temporal regions during auditory word presentation and verb generation. Cerebr. Cortex *6*, 1–10.

Giard, M.H., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. J. Cogn. Neurosci. *11*, 473–490.

Gold, J., Bennett, P.J., and Sekuler, A.B. (1999). Signal but not noise changes with perceptual learning. Nature *402*, 176–178.

Gray, C.M., and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc. Nat. Acad. Sci. USA *86*, 1698–1702.

Hämäläinen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., and Lounasmaa, O.V. (1993). Magnetoencephalography: theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev. Mod. Physics *65*, 413–497.

Howard, D., Patterson, K., Wise, R., Brown, W.D., Friston, K., Weiller, C., and Frackowiak, R. (1992). The cortical localization of the lexicons. Brain *115*, 1769–1782.

Hurlbert, A. (2000). Visual perception: learning to see through noise. Curr. Biol. *10*, R231–R233.

Iacoboni, M., Woods, R.P., and Mazziotta, J.C. (1998). Bimodal (auditory and visual) left frontoparietal circuitry for sensorimotor integration and sensorimotor learning. Brain *121*, 2135–2143.

Kadunce, D.C., Vaughan, J.W., Wallace, M.T., Benedek, G., and Stein, B.E. (1997). Mechanisms of within- and cross-modality suppression in the superior colliculus. J. Neurophysiol. *78*, 2834–2847.

Kok, A., and de Jong, H.L. (1980). The effect of repetition of infrequent familiar and infamiliar visual patterns on components of the event-related brain potential. Biol. Psychol. *10*, 167–188.

Livingstone, M., and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science *240*, 740–749.

Martin, A., Wiggs, C.L., Ungerleider, L.G., and Haxby, J.V. (1996). Neural correlates of category-specific knowledge. Nature *379*, 649–652.

Matsuura, K., and Okabe, U. (1995). Selective minimum-norm solution of the biomagnetic inverse problem. IEEE Trans. Biomed. Eng. *42*, 608–615.

McCarthy, R.A., and Warrington, E.K. (1988). Evidence for modality-specific meaning systems in the brain. Nature *334*, 428–430.

McGrath, M., and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. J. Acoust. Soc. Am. *77*, 678–685.

Mesulam, M.-M. (1998). From sensation to cognition. Brain *121*, 1013–1052.

Miller, J. (1986). Timecourse of coactivation in bimodal divided attention. Percept. Psychophys. *40*, 331–343.

Morrell, L.K. (1968). Sensory interaction: evoked potential observations in man. Exp. Brain Res. *6*, 146–155.

Ojemann, G. (1992). Localization of language in frontal cortex. Adv. Neurol. *57*, 361–368.

Ojemann, G., Ojemann, J., Lettich, E., and Berger, M. (1989). Cortical language localization in left, dominant hemisphere. An electrical stimulation mapping investigation in 117 patients. J. Neurosurg. *71*, 316–326.

Pandya, D.P., and Yeterian, E.H. (1985). Architecture and connections of cortical association areas. In Cerebral Cortex: Association and Auditory Cortices, Volume 4, A. Peters and E.G. Jones, eds. (New York: Plenum Press), pp. 3–61.

Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S.F., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., et al. (2000). A cultural effect on brain function. Nat. Neurosci. *3*, 91–96.

Paus, T. (1996). Location and function of the human frontal eye-field: a selective review. Neuropsychol. *34*, 475–483.

Peck, C.K., Baro, J.A., and Warder, S.M. (1995). Effects of eye position on saccadic eye movements and on the neuronal responses to auditory and visual stimuli in cat superior colliculus. Exp. Brain. Res. *103*, 227–242.

Price, C.J., Wise, R.J.S., and Frackowiak, R.S.J. (1996). Demonstrating the implicit processing of visually presented words and pseudowords. Cereb. Cortex *6*, 62–70.

Price, C.J., Moore, C.J., Humphreys, G.W., and Wise, R.J.S. (1997). Segregating semantic from phonological processes during reading. J. Cogn. Neurosci. *9*, 727–733.

Raab, D.H. (1962). Statistical facilitation of simple reaction times. Trans. NY Acad. Sci. *24*, 574–590.

Raichle, M.E., Fiez, J.A., Videen, T.O., MacLeod, A.M., Pardo, J.V., Fox, P.T., and Petersen, S.E. (1994). Practice-related changes in human brain functional anatomy during nonmotor learning. Cereb. Cortex *4*, 8–26.

Raij, T.A. (1999). Patterns of brain activity during visual imagery of letters. J. Cogn. Neurosci. *11*, 282–299.

Roelfsema, P.R., Engel, A.K., König, P., and Singer, W. (1996). The role of neuronal synchronization in response selection: a biologically plausible theory of structured representations in the visual cortex. J. Cogn. Neurosci. *8*, 603–625.

Roland, P.E., and Zilles, K. (1996). The developing European computerized human brain database for all imaging modalities. Neuroimage *4*, S39–S47.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.-T., and Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci. Lett. *127*, 141–145.

Schröger, E., and Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. Psychophysiol. *35*, 755–759.

Seltzer, B., Cola, M.G., Gutierrez, C., Massee, M., Weldon, C., and Cusick, C.G. (1996). Overlapping and nonoverlapping cortical projections to cortex of the superior temporal sulcus in the rhesus monkey: double anterograde studies. J. Comp. Neurol. *370*, 173–190.

Sergent, J., Zuck, E., Levesque, M., and MacDonald, B. (1992). Positron emission tomography study of letter and object processing: empirical findings and methodological considerations. Cereb. Cortex *2*, 68–80.

Squires, N.K., Donchin, E., Squires, K.C., and Grossberg, S. (1977). Bisensory stimulation: inferring decision-related processes from P300 component. J. Exp. Psychol. Hum. Percept. Perf. *3*, 299–315.

Stein, B.E. (1998). Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. Exp. Brain. Res. *123*, 124–135.

Stein, B.E., and Meredith, M.A. (1993). The Merging of the Senses (Cambridge, MA: MIT Press).

Stein, B.E., and Wallace, M.T. (1996). Comparisons of cross-modality integration in midbrain and cortex. Progr. Brain Res. *112*, 289–299.

Stein, B.E., Huneycutt, W.S., and Meredith, M.A. (1988). Neurons and behavior: the same rules of multisensory integration apply. Brain Res. *448*, 355–358.

Talairach, J., and Tournoux, P. (1988). Co-Planar Stereotaxic Atlas of the Human Brain (New York: Thieme).

Thompson, R.F., Johnson, R.H., and Hoopes, J.J. (1962). Organization of auditory, somatic sensory, and visual projection to association fields of cerebral cortex in the cat. J. Neurophysiol. *26*, 343–364.

Uutela, K., Hämäläinen, M., and Somersalo, E. (1999). Visualization of magnetoencephalographic data using minimum current estimates. Neuroimage *10*, 173–180.

Wallace, M.T., and Stein, B.E. (1994). Cross-modal synthesis in the midbrain depends on input from cortex. J. Neurophysiol. *71*, 429–432.

Wallace, M.T., Wilkinson, L.K., and Stein, B.E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. J. Neurophysiol. *76*, 1246–1266.

Warrington, E.K., and Taylor, A. (1973). Contribution of the right parietal lobe to object recognition. Cortex *9*, 152–164.

Wiggs, C.L., and Martin, A. (1998). Properties and mechanisms of perceptual priming. Curr. Opin. Neurobiol. *8*, 227–233.

Yaka, R., Yinon, U., and Wollberg, Z. (1999). Auditory activation of cortical visual areas in cats after early visual deprivation. Eur. J. Neurosci. *11*, 1301–1312.